PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : H04L 12/28, 12/43, 12/56, H04J 3/16, 3/22 — A1

(11) International Publication Number: WO 99/63712

(43) International Publication Date: 9 December 1999 (09.12.99)

(21) International Application Number: PCT/US99/10592

(22) International Filing Date: 13 May 1999 (13.05.99)

(30) Priority Data:
09/087,064 29 May 1998 (29.05.98) US

(71) Applicant: CABLETRON SYSTEMS, INC. [US/US]; 35 Industrial Way, Rochester, NH 03866 (US).

(72) Inventor: AATRESH, Deepak, J.; 197 Cirrus Avenue, Sunnyvale, CA 94087 (US).

(74) Agent: WILSON, Mark; Law Offices of Mark Wilson, PMB: 348, 2530 Berryessa Road, San Jose, CA 95132 (US).

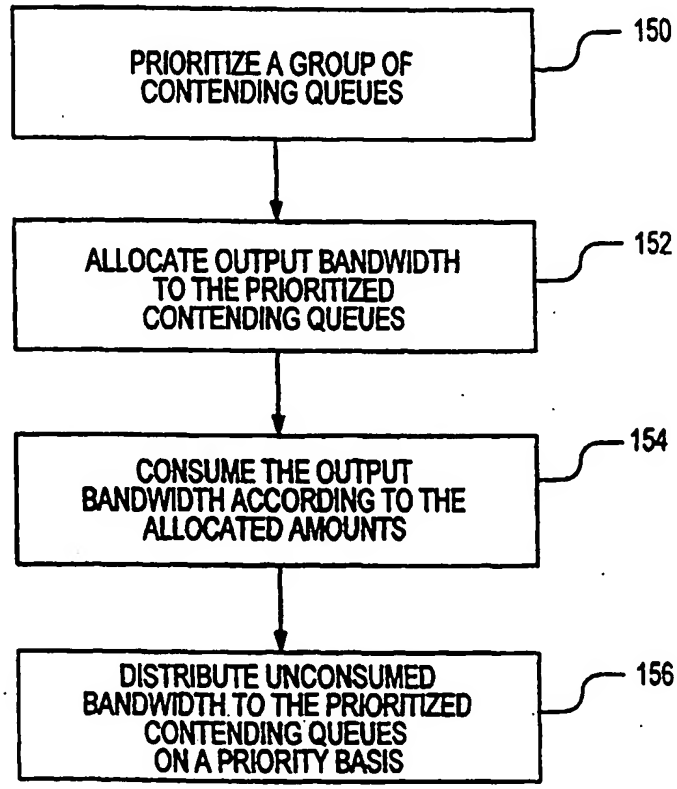
(81) Designated States: AU, CA, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: METHOD AND APPARATUS FOR FORWARDING PACKETS FROM A PLURALITY OF CONTENDING QUEUES TO AN OUTPUT

(57) Abstract

A method for prioritizing queues (150) wherein output bandwidth is allocated (152), output bandwidth is consumed (154), and unconsumed bandwidth is distributed to the prioritized contending queues (156).



METHOD AND APPARATUS FOR FORWARDING PACKETS FROM A PLURALITY OF CONTENDING QUEUES TO AN OUTPUT

TECHNICAL FIELD

The invention relates generally to a high bandwidth multiport switch, for instance as used in gigabit ethernet networks. More particularly, the invention relates to the buffering of packets within a switch and maintaining Quality of Service (QoS) through the switch.

BACKGROUND OF THE INVENTION

Networks are used to transfer voice, video, and data between various network devices. Network devices such as switches are located within networks to direct the transfer of network traffic between the various network devices. Network traffic is typically bursty in nature and in order to compensate for network traffic bursts, memory buffers have been incorporated into switch designs. Memory buffers allow a switch to temporarily store packets when an incoming data rate is higher than an outgoing data rate. When more than one buffer has packets that are contending for the same output, some packets are required to wait in the buffers and some mechanism is needed to determine how packet contention is resolved.

In order to resolve contention and provide a higher QoS to a network switch, two different buffer management schemes, priority queuing and weighted fair queuing, are known. In the priority queuing scheme, contending queues are assigned different priorities and packets are forwarded from the queues in strict priority order. For example, referring to Fig. 1, four queues A, B, C, and D buffer packets that are contending for the same output 20 that has a finite bandwidth. To resolve the contention, the queues are assigned different priorities such as first, second, third, and fourth priority. Packets are similarly prioritized and queued according to their priority. The queued packets are forwarded to the output in priority order such that all packets from higher priority queues are forwarded before any packets from lower priority queues. While priority queuing works well when there is contention between queues, the higher priority packets are forwarded at the expense of lower priority packets. When higher priority packets consume the majority of the finite output bandwidth, the lower priority packets are starved

easier to develop and more flexible to update than logic embedded into application-specific integrated circuits (ASICs), operations performed in software typically take more time and have greater overhead relative to logic that is embedded into ASICs. As bandwidth requirements for networks increase, the speed with which contending packets are released from buffers is of greater concern.

In view of the above-stated disadvantages of the prior art, what is needed is a way to avoid the starvation of prioritized queues while logically distributing unused bandwidth at speeds required by, for example, gigabit ethernet networks.

SUMMARY OF THE INVENTION

A method and apparatus for forwarding packets from contending queues to an output having a finite bandwidth involve prioritizing the contending queues, allocating shares of the output bandwidth to the prioritized queues, forwarding prioritized packets in accordance with the allocated proportions, and then distributing any unconsumed bandwidth to the prioritized queues on a priority basis. In a preferred embodiment, the unconsumed bandwidth is offered to the queues in priority order from the highest priority queue to the lowest priority queue. Further, in the preferred embodiment, the logic for performing queue management is embedded in an application specific integrated circuit.

The method and apparatus of the invention are preferably implemented in a high bandwidth multiport switch in which packets are buffered before being output from the switch. The architecture of the multiport switch includes data links which are connected to input/output controllers which are connected to a switch fabric. The data links provide the data paths between other devices in the network and the multiport switch. There may be multiple data links of varying types and capacities connected to the multiport switch. Preferably, the data links are twisted pair wires and optical fibers that carry variable-length packets at 10, 100, and 1,000 Mbps according to the ethernet protocol.

The input/output controllers are connected between the data links and the switch fabric to provide packet control between the data links and the switch fabric. Packet control functions include transmitting and receiving packets from and to the data links as well as buffering incoming and/or outgoing packets to accommodate fluctuations in network traffic rates.

because the quality of the video conference is negatively affected when packets are delayed while an e-mail transfer is more tolerant of packet delays. Four queues that track the established packet priority categories are associated with each port of an input/output controller. The four queues track the four priority categories and both the packet and queue priorities are defined, for description purposes, as control (CTRL), high (HI), medium (MED), and low (LO), where control is the highest priority and low is the lowest priority. The queue scheme is not limited to four queues and can be scaled up or down.

The input/output controllers receive packets from the switch fabric, and if the packets are not immediately forwarded to a data link, the packets are buffered in memory and the appropriate queue within the queue manager is updated with information related to the buffered packet. When priority queuing is the only queue management scheme being utilized, all packets of a higher priority are forwarded from the prioritized queues before any packets of a lower priority. For example, any packets in the control queue will be forwarded to an associated data link before packets from the high, medium, or low queues. In the queue management scheme of the invention, weighted fair queuing is implemented in conjunction with priority queuing to better utilize the finite bandwidth of an output.

In operation, each one of the four priority queues is allocated a respective share of the total output data link bandwidth to be consumed over a designated period of time. During each designated time period, packets are released from the queues to the associated data link in priority order according to their allocated shares such that the allocated packets in the control queue are released before the allocated packets in the high queue and so on. The key to the queue management scheme of the invention is that when a queue does not consume its entire allocated bandwidth during a designated time interval, the excess bandwidth is allocated to the other queues on a priority basis. That is, excess bandwidth is first offered to the control queue, and if the control queue has enough waiting packets then the entire excess bandwidth is consumed by the control queue. On the other hand, if the control queue does not consume all of the excess bandwidth then the excess bandwidth is offered to the next lower priority queue and so on until all of the available output bandwidth is consumed or until no packets remain in the output buffer.

One of the main operating considerations of the weighted priority queuing scheme of the invention is the balancing between latency

Fig. 3 is a depiction of the basic switch architecture of the preferred embodiment of the invention.

Fig. 4 is an expanded depiction of a preferred architecture of an input/output controller with one data link in accordance with the invention.

Fig. 5 is an expanded depiction of a preferred architecture of an input/output controller with two data links in accordance with the invention.

Fig. 6 is a depiction of the weighted priority queuing scheme in accordance with the invention.

Fig. 7 is a process flow diagram for forwarding packets in accordance with the invention.

Fig. 8 is a depiction of the basic hardware architecture for performing weighted priority queuing in accordance with the invention.

Fig. 9 is a depiction of internal register values used for allocating bandwidth between four priority queues.

DETAILED DESCRIPTION

Fig. 3 is a depiction of the basic architecture of a switch 38 for forwarding variable-length packets that includes the preferred embodiment of the invention. Although a four-channel switch is shown for description purposes, the switch may have fewer but preferably has more channels. The preferred architecture includes data links 52, 54, 56, and 58 which are connected to input/output (I/O) controllers 42, 44, 46, and 48 which are connected to a switch fabric 40.

The data links 52-58 connected to the I/O controllers 42-48 provide the medium for transferring packets of data into and out of the switch 38. In a preferred embodiment, the number of data links connected to each I/O controller is based on the bandwidth capacity of the data link. For example, in Fig. 3 the single and double data links 52, 54, and 56 represent 1,000 Megabits per second (Mbps) connections and the eight data links 58 represent ten and/or 100 Mbps connections, although these connection bandwidths can be larger or smaller and the number of data links per I/O controller can be larger or smaller. In addition, the physical makeup of the data links is preferably twisted pair wires and/or single mode optical fibers, although other data links such as coaxial cable, multimode optical fiber, infrared, and/or radio frequency links, are possible.

The I/O controllers 42-48 are connected directly to the data links 52-58 and are connected to the switch fabric 40 by input

controller and in the preferred embodiment, the output buffer is located next to the I/O controller and not integrated onto the same circuit as the I/O controller, although this is not critical to the invention.

In the preferred embodiment, packets are prioritized into four categories based on certain characteristics of the packets. The packet characteristics of interest to the prioritization may include the source and/or destination of the packet, the type of information carried in the packet, or the age of the packet. For example, a packet carrying video conferencing data may have a higher priority than a packet carrying e-mail, because the quality of the video conference is negatively affected when packets are delayed while an e-mail transfer is more tolerant of packet delays. The four queues 84-90 in Fig. 4 track the established packet priority categories and the four priority categories for both the packets and the queues are defined, for description purposes, as control (CTRL), high (HI), medium (MED), and low (LO) where control is the highest priority and low is the lowest priority. Each of the four prioritized queues is depicted as containing eight registers relating to eight packets, although the exact number is not critical to the invention. The capacity of the queues may also be changed through a programming interface. The I/O controller of Fig. 4 receives packets from the switch fabric through the input connection 72, and if the packet is not immediately forwarded for output to the data link 52, the packet is buffered in memory and the appropriate queue within the queue manager is updated with information related to the buffered packet. In the preferred embodiment, the designation of packet and/or queue priorities is changeable during operation, or on the fly, through a programming interface.

Fig. 5 is an expanded depiction of an I/O controller 46 of Fig. 3 that has two data links 56 and 56a. The I/O controller connected to the two data links includes separate output queue managers 94 and 96 for each of the data links. As with the I/O controller of Fig. 4, the two queue managers of Fig. 5 have four priority queues 100, 102, 104, 106 and 110, 112, 114, 116, respectively, where each queue contains eight registers related to eight packets. The output queue managers are related on a one-to-one basis to the two data links connected to the I/O controller and both of the output queue managers receive packets from the switch fabric through the same input connection 76. Although not depicted in an expanded view, an I/O controller 48 from Fig. 3 that has eight data links 58 has eight output queue managers corresponding on a one-to-one basis to the eight data links.

assume that the control priority queue has more than enough packets to consume 10% of the data link bandwidth and so the 10% of the bandwidth is consumed. Next, the high priority queue 132 is offered 30% of the bandwidth and, for example purposes, assume that the high priority queue has more than enough packets to consume its 30% of the bandwidth and so the 30% of bandwidth is consumed. Next, the medium priority queue 134 is offered 15% of the bandwidth and, for example purposes, assume that the medium priority queue has no packets waiting to be forwarded. In order to avoid wasting the allocated bandwidth and to maximize the utilization of the available bandwidth of the data link, the 15% of the data link bandwidth that is allocated to the medium priority queue for the present time interval is offered to the higher priority queues in priority order. That is, the 15% of allocated data link bandwidth from the medium priority queue is first offered entirely to the control priority queue. If the control priority queue has enough packets waiting, it will consume all of the excess bandwidth, or if the control priority queue does not consume all of the excess bandwidth then the remaining excess bandwidth is offered to the high priority queue. Further, if the high priority queue does not consume the excess bandwidth, then the excess bandwidth is offered to the low priority queue 136. Under this prioritized distribution of excess bandwidth, excess bandwidth will never be wasted as long as there are buffered packets.

Allocated bandwidth in the weighted priority queuing scheme can be adjusted through the programming interface to specify a certain QoS and to create different traffic patterns. For example, referring to Example 2 of Fig. 6, the control priority queue 130 is allocated 30% of the total data link bandwidth, the high priority queue 132 is allocated 0% of the total data link bandwidth, the medium priority queue 134 is allocated 50% of the total data link bandwidth, and the low priority queue 136 is allocated 20% of the total data link bandwidth. When a queue, such as the high priority queue, is allocated 0% of the bandwidth, in effect the queue with 0% allocated bandwidth and the next higher priority queue share excess bandwidth according to a pure priority scheme. In Example 2, any excess bandwidth from the control priority queue is offered first to the high priority queue even though the high priority queue is allocated 0% of the bandwidth. In Example 3, any excess bandwidth offered to the high priority queue is shared between the high priority queue and the medium priority queue according to a pure priority scheme even though the medium priority queue is allocated 0% of the bandwidth.

When balancing latency and error to provide a specified QoS, a relatively long bandwidth allocation cycle time will create high packet latency but a low error rate. If latency is too high; queues may be practically starved between cycle times and packets may begin to be dropped. On the other hand, a relatively short bandwidth allocation cycle time will create low packet latency but a high error rate as more packets exceed the relatively short queue-specific time intervals that are allocated for packet transmission. Implementing weighted priority queuing in a 100 Mbps ethernet network where packets sizes range from 64 bytes to 1,500 bytes, a bandwidth allocation cycle time, or latency, of 1.28 ms has an error of approximately 9.38% and a latency of 163.84 ms has an error of approximately 0.7%. The error distribution scales with other data rates.

Fig. 7 is a process flow diagram for forwarding packets in accordance with a preferred embodiment of the invention. In a first step 150, a group of contending queues are prioritized relative to one another from a highest priority queue to a lowest priority queue. In a next step 152, portions of the bandwidth of an output are allocated to the group of prioritized queues. In a next step 154, the bandwidth of the output is consumed by packets from the prioritized queues according to the portions of bandwidth that are allocated to each queue. In a next step 156, bandwidth that is not consumed by the queues according to the allocated portion is distributed to the queues on a priority basis. In the preferred embodiment, the unconsumed bandwidth is first offered to the highest priority queue before any other queues and the unconsumed bandwidth is only offered to the lowest priority after the bandwidth has been offered to all of the higher priority queues.

Fig. 8 is a depiction of the basic architecture of a hardware implementation of the preferred weighted priority queuing scheme. The preferred hardware includes four prioritized queues 180, 182, 184, and 186, a multiplexer 190, priority logic 194, and weighted fair queuing logic 198 embedded in an ASIC. The weighted fair queuing logic includes registers 200, 202, 204, 206, and 208, a counter 210, and comparators 220, 222, 224, and 226. The registers are used to establish the bandwidth allocation proportions of the queues. The counter is decremented from a maximum value at regular time intervals as dictated by a system clock. The comparators compare the counter value to the register values to determine which priority queue should have access to the output 230. When the counter value drops below one of the register values, the respective comparator is tripped and a select signal is sent from the comparator to the priority logic. The select

WHAT IS CLAIMED IS:

1. A method for forwarding packets from a plurality of contending queues to an output having a finite bandwidth comprising the steps of:

prioritizing said plurality of queues such that each of said queues has a priority relative to the other queues of said plurality of queues, thereby defining a prioritization range that includes a highest priority queue and a lowest priority queue;

allocating a share of said bandwidth of said output to each of said queues;

consuming at least a portion of said bandwidth of said output with packets from said queues according to said allocated shares, leaving an unconsumed portion when at least one of said queues does not exhaust the allocated share of said at least one queue; and

distributing said unconsumed portion of said bandwidth to said queues according to said prioritization range.

2. The method of claim 1 wherein said step of prioritizing said plurality of queues includes a step of prioritizing each of said queues with a different priority relative to the other queues of said queues.

3. The method of claim 2 further comprising a step of prioritizing said packets into a prioritization range that directly relates to said prioritization range of said queues.

4. The method of claim 2 wherein said step of distributing said unconsumed portion of said bandwidth includes steps of:

offering said unconsumed portion to said highest priority queue before any other of said queues; and

offering said unconsumed portion to said lowest priority queue only after said unconsumed portion has been offered to higher priority queues of said queues.

9. An application-specific integrated circuit (ASIC) having a plurality of queues related to packets contending for the same output, where said output has a bandwidth capacity comprising:

means for identifying a priority order among said plurality of queues wherein one queue has a highest priority among said queues and a different queue has a lowest priority among said queues;

means for allocating a percentage of said bandwidth capacity of said output to each of said queues;

means, formed in circuitry that is specific to said queue management, for forwarding packets to said output according to said allocated percentages; and

means, formed in circuitry that is specific to queue management, for distributing unused bandwidth capacity to said plurality of queues in priority order from said queue with said highest priority to said queue with said lowest priority.

10. The ASIC of claim 9 wherein said means for allocating is formed in circuitry that is specific to said queue management.

11. The ASIC of claim 9 wherein said means for allocating includes run-time programmable registers that are set to register values that define time intervals.

12. The ASIC of claim 11 wherein said means for forwarding includes a clock counter that generates counter values.

13. The ASIC of claim 12 wherein said means for forwarding packets includes a plurality of comparators, wherein each comparator has a first input for receiving one of said counter values and a second input for receiving one of said register values.

16. A method of forwarding contending variable-length packets from output queues to an output port of a multiport switch comprising steps of:

prioritizing said output queues such that each output queue has a different priority relative to the other of said contending queues;

storing said contending variable-length packets in output queues of said prioritized output queues;

allocating bandwidth of said output port among said prioritized output queues such that said prioritized output queues have allocated shares of said bandwidth;

forwarding contending variable-length packets from said prioritized output queues to said output ports according to said bandwidth allocations;

monitoring said forwarding of contending variable-length packets to detect bandwidth allocations in excess of bandwidth consumed by said contending variable-length packets in said prioritized output queues; and

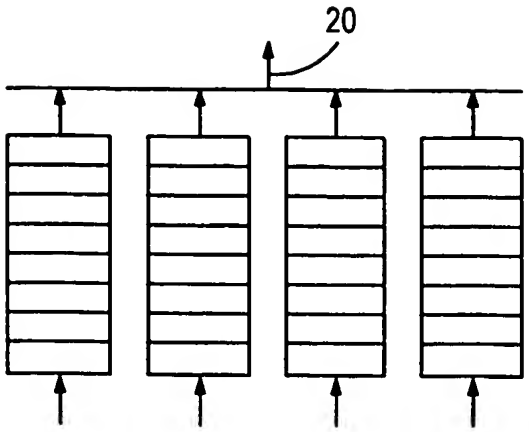
offering excess bandwidth allocations that are detected by said monitoring to a highest priority output queue of said prioritized output queues that has a variable-length packet to be forwarded to said output port of said multiport switch.

17. The method of claim 16 wherein said step of storing includes steps of:

prioritizing said contending variable-length packets; and

storing said prioritized contending variable-length packets in similarly prioritized queues of said prioritized output queues.

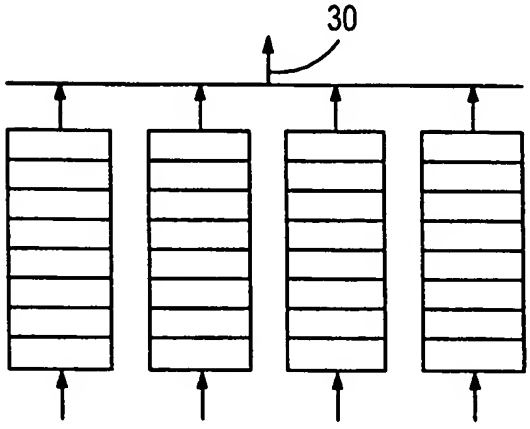
18. The method of claim 16 wherein said step of allocating bandwidth includes a step of setting register values in an application specific integrated circuit on a queue specific basis while said multiport switch is forwarding packets.



QUEUE NAME	A	B	C	D
QUEUE PRIORITY	1ST	2ND	3RD	4TH

PRIORITY QUEUEING

FIG. 1
(PRIOR ART)



QUEUE NAME	A	B	C	D
QUEUE WEIGHT	20%	25%	15%	40%

WEIGHTED FAIR QUEUEING

FIG. 2
(PRIOR ART)

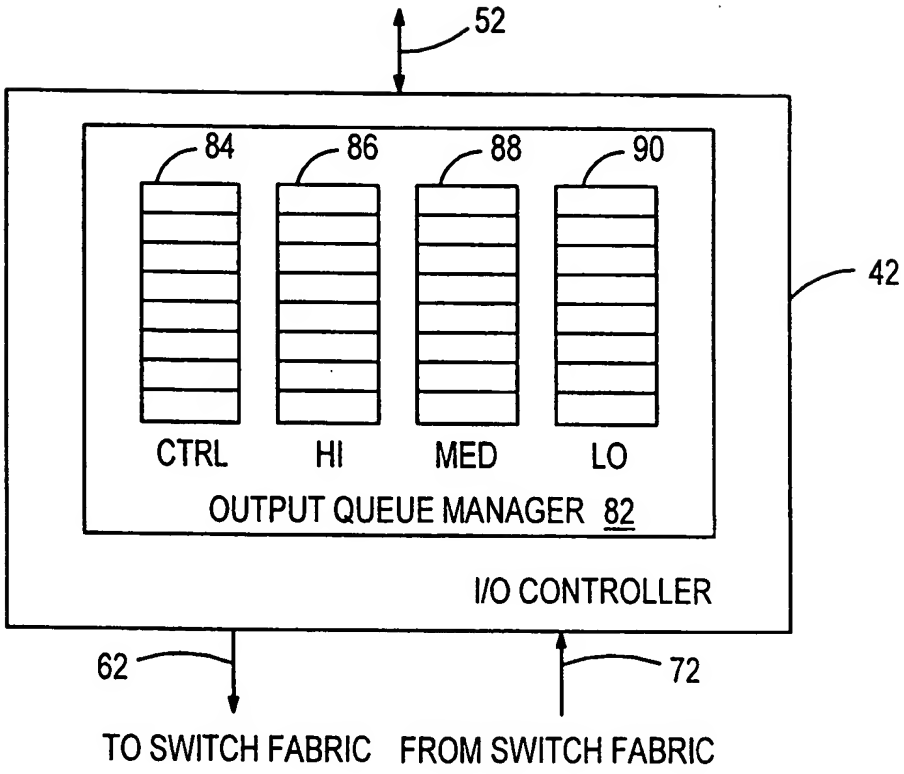
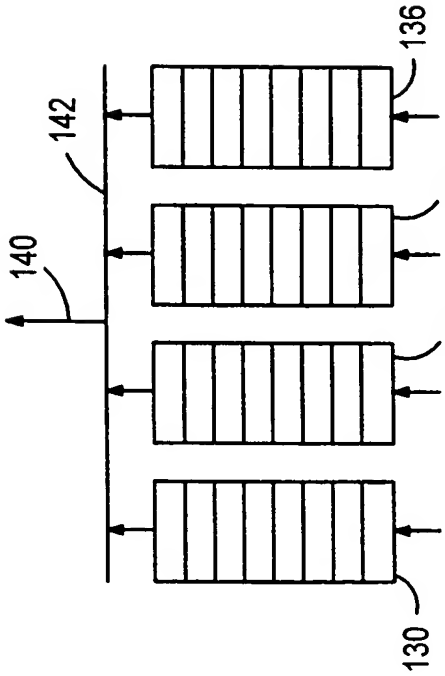

52
84
86
88
90
42
CTRL
HI
MED
LO
OUTPUT QUEUE MANAGER 82
I/O CONTROLLER
62
72
TO SWITCH FABRIC
FROM SWITCH FABRIC

FIG. 4



PACKET PRIORITY	CTRL	HI	MED	LO
BANDWIDTH ALLOCATION (%)	10	30	15	45
	30	0	50	20
	30	50	0	20

EXAMPLE 1
EXAMPLE 2
EXAMPLE 3

FIG. 6

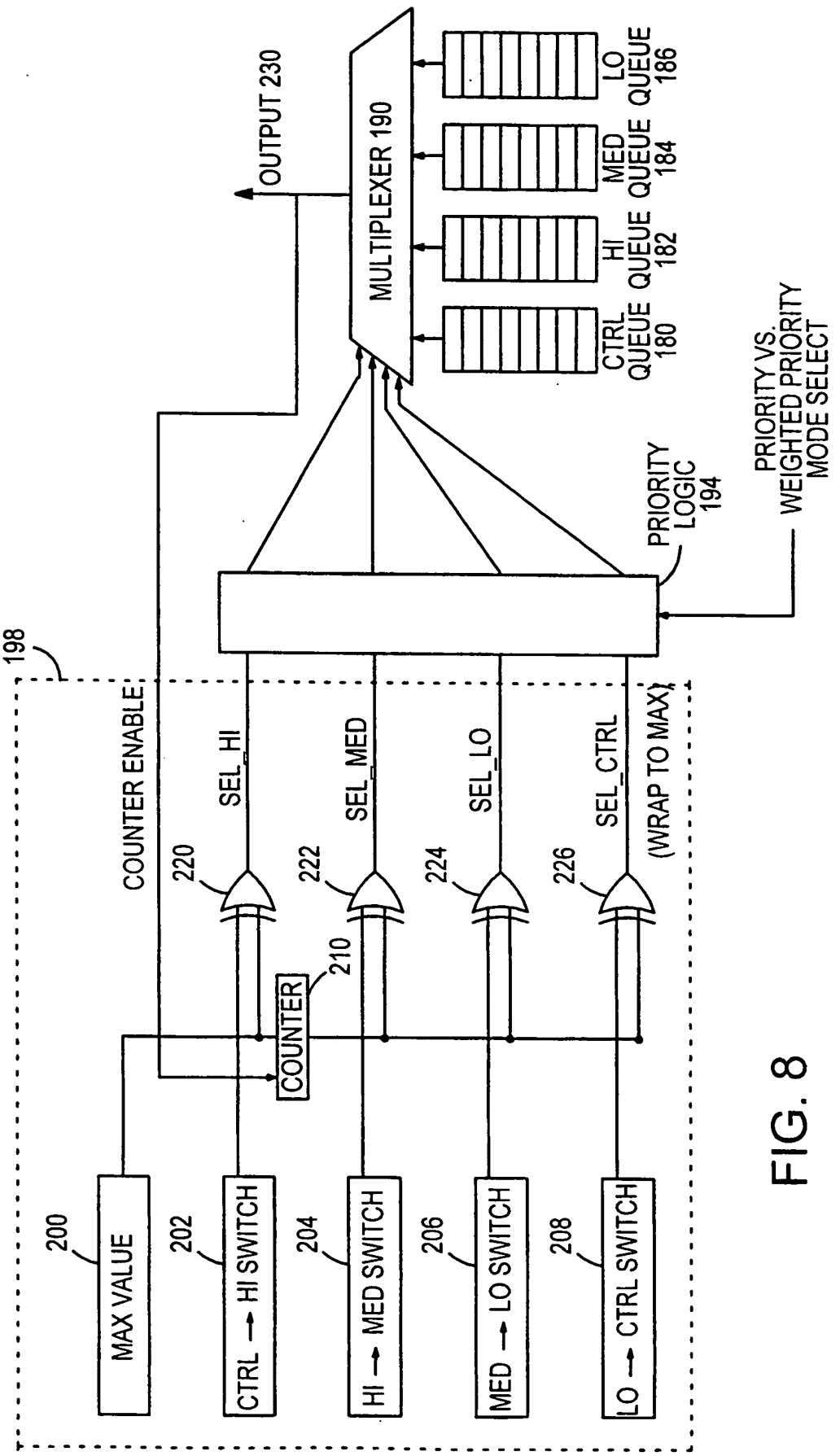

198
COUNTER ENABLE
SEL_HI
SEL_MED
SEL_LO
SEL_CTRL
(WRAP TO MAX)
220
222
224
226
210
COUNTER
200
MAX VALUE
202
CTRL → HI SWITCH
204
HI → MED SWITCH
206
MED → LO SWITCH
208
LO → CTRL SWITCH
OUTPUT 230
MULTIPLEXER 190
CTRL QUEUE 180
HI QUEUE 182
MED QUEUE 184
LO QUEUE 186
PRIORITY LOGIC 194
PRIORITY VS. WEIGHTED PRIORITY MODE SELECT

FIG. 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/10592

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :H04L 12/28, 12/43, 12/56; H04J 3/16, 3/22
US CL : 370/412-418, 461, 468

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 370/412-418, 461, 468

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

NONE

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS
search terms: asic, priorit?, queu?, register#, counter#, bandwidth, switch?

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,748,629 A (CALDARA et al.) 05 MAY 1998, Figs. 1, 4, 6-8, 11, summary, col. 4, lines 7-30; col. 7, lines 19-26; col. 8, lines 17-31; col. 10, lines 13-29; col. 13, lines 1-9	1-20

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 17 JUNE 1999	Date of mailing of the international search report 16 JUL 1999
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer DAVID R. VINCENT Telephone No. (703) 305-4957